基于改进引力搜索算法的 K-means 聚类 *

魏康园 [a, b]，何　庆 [a, b†]，徐钦帅 [a, b]

(贵州大学 a. 大数据与信息工程学院; b. 贵州省公共大数据重点实验室，贵阳 550025)

摘　要：针对 K-means 算法的聚类结果极易受到聚类中心的影响而陷入局部最优解的问题，提出一种基于改进引力搜索的 K-means 聚类算法。首先引入自适应概念，对引力系数衰减因子进行控制，提高算法的全局探索能力和局部开发能力；然后，引入免疫克隆选择机制，以便算法能够有效跳出局部最优，并通过对 12 个基准测试函数的实验验证改进引力搜索算法的有效性和优越性；最后，通过结合改进的引力搜索算法和 K-means 算法，提出一种新的聚类算法 A2F-GSA-Kmeans，并在 6 个测试数据集上的实验表明，该算法具有较好的聚类质量。

关键词：K-means 算法；引力搜索算法；引力系数衰减因子；免疫克隆选择算法

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2018.06.0310

Novel K-means clustering algorithm based on improved gravitational search algorithm

Wei Kangyuan[a, b], He Qing[a, b†], Xu Qinshuai[a, b]

(*a. College of Big Data & Information Engineering*, *b. Guizhou Provincial Key Laboratory of Public Big Data Guizhou University*, *Guiyang* 550025, *China*)

Abstract: In order to solve the problem that the clustering result of K-means algorithm gets affected by the initial cluster centers easily, this paper proposed a novel K-means clustering algorithm based on improved gravitational search algorithm. Firstly, it enhanced the global exploration and local exploitation capability of the algorithm with the introduction of adaptive concept to control the attenuation factor of gravitational constant. Then, by introducing immune clonal selection algorithm to make the algorithm jump out of the local optimum efficiently. The experimental results on twelve test functions prove the effectiveness and superiority of the improved GSA. Finally, by combining the improved GSA with K-means algorithm, this paper proposed a new clustering algorithm called A2F-GSA-Kmeans. The experimental results on six test datasets show that the algorithm has better clustering quality.

Key words: K-means clustering algorithm; gravitational search algorithm; attenuation factor of gravitational constant; immune clonal selection algorithm

0　引言

聚类是分析数据的重要方法，其作用是将多个抽象对象按照相应的标准分成由相似的对象组成的多个类过程，已经被广泛应用于许多领域，其中 K-means 算法在处理大量数据时具有简单高效的优点，已经被广泛应用，但其聚类结果极易受到聚类中心的影响，导致陷入局部最优解错误!未找到引用源。，并且要求用户指定聚类数量，然而不同的聚类数将会得到不同的聚类结果，直接影响算法的效率。因此，算法本身可以获得最优数目的聚类是非常重要的。

群智能算法因其强大的全局搜索能力，已经被越来越多的学者应用到聚类问题中，来弥补传统聚类算法的缺陷。例如杨菊蜻等错误!未找到引用源。通过对蝙蝠优化算法（BA）的位置和速度更新方式进行优化，同时引入非线性惯性权重和 limit 阈值思想，提高了算法的收敛性能，并将改进算法与 K-means 结合，提出了一种基于改进 BA 算法的 K-means 算法，取得了较好的聚类效果；于佐军等人错误!未找到引用源。通过引入算术交叉操作改进人工蜂群算法中引领蜂和跟随蜂的搜索模式，并结合 K-means 算法，提出一种聚类算法来自动寻找最优的聚类数。

引力搜索算法错误!未找到引用源。（Gravitational Search Algorithms，GSA）是 Esmat Rashedi 教授等于 2009 年提出的一种新的群智能优化算法，该算法通过模拟物理学中万有引力来搜索全局最

收稿日期：2018-06-20；**修回日期：**2018-07-27　　**基金项目：**贵州省公共大数据重点实验室开放课题（2017BDKFJJ004）；贵州省教育厅青年科技人才成长项目（黔科合 KY 字［2016］124）；贵州大学培育项目（黔科合平台人才［2017］5788）

作者简介：魏康园（1991-），女，陕西渭南人，硕士研究生，主要研究方向为数据挖掘、进化计算；何庆（1982-），男（通信作者），贵州贵阳人，副教授，博士，主要研究方向为大数据应用、人工智能（qhe@gzu.edu.cn）；徐钦帅（1994-），男，山东枣庄人，硕士研究生，主要研究方向为机器学习、进化计算.

优解。研究发现，在对基准测试函数进行优化时，经典 GSA 算法的优化精度与收敛速度均明显优于粒子群优化算法（particle swarm optimization，PSO）和遗传算法（genetic algorithm，GA）等优化算法错误!未找到引用源。。但是，与其他元启发式算法类似，经典 GSA 算法同样存在早熟收敛、易陷入局部极值等缺陷，基于此，近年来国内外学者实现了许多有效的改进 GSA 算法，如 Liu 等人错误!未找到引用源。利用混沌映射优化 GSA 算法中粒子的位置初始化，同时将自适应递减惯性权重系数引入到位置更新公式中，提出并实现了 AC-GSA 算法，并在经典基准测试函数的测试和优化最小二乘支持向量机超参数上都取得了良好的结果；Sun 等人错误!未找到引用源。基于粒子个体的异质性，利用粒子个体最优值和全局最优值对 GSA 算法的 Kbest 和速度更新方式进行改进，提出了 LIGSA 算法，使得粒子能够学习 K 个近邻粒子而充分开发搜索空间并有效防止早熟收敛，同时全局最优值的引导可加速算法收敛速度；Mirjalili 等人错误!未找到引用源。基于 GSA 算法中引力系数 G 对于平衡算法全局探索能力与局部开发能力的重要性，利用混沌映射对引力系数进行改进，并验证了其能够跳出局部最优实现更高寻优精度的有效性。然而，尽管已有研究对经典 GSA 算法的寻优效果有所提高，且大多集中于结合 PSO 算法对 GSA 算法的速度和位置更新方式的改进错误!未找到引用源。8~错误!未找到引用源。，但对于实现算法探索与开发能力的有效平衡和解决其早熟收敛问题仍需深入研究。

综上所示，本文针对引力搜索算法易陷入局部最优、发生早熟收敛现象等问题，首先对引力搜索算法中的引力系数作出改进，保留算法较高精度和收敛速度的同时，提高算法的全局探索能力和局部最优能力；然后再引入免疫克隆选择机制，改善算法早熟收敛现象；最后将改进的 GSA 算法应用到 K-means 聚类算法中，通过调节 K 值大小，利用聚类评价函数来获取最佳的聚类数。

1 相关工作

1.1 K-means 算法

K-means 算法是一种基于划分的聚类算法，该算法首先随机选取样本空间中的 K 个数据点作为聚类中心，然后通过计算样本中其他数据点与中心点的欧式距离大小，来对数据进行划分。该算法的流程如图 1 所示。

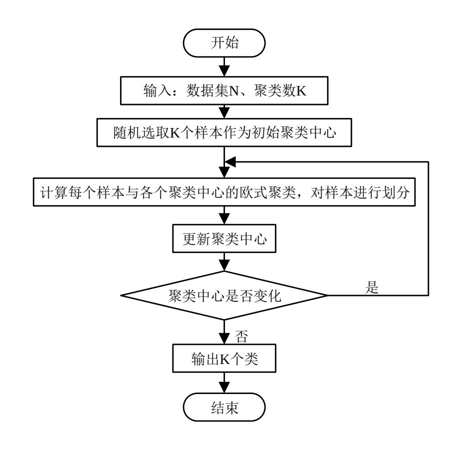


图 1 K-means 聚类算法流程图

对于欧氏距离的数据，本文使用均方误差 MSE 作为聚类的目标函数，且 MSE 的值越小表示聚类效果越好。定义为

$$MSE=\frac{1}{n}\sum_{j=1}^{k}\sum_{y_i\in C_j}\left\|y_i-z_j\right\|^2 \tag{1}$$

其中，z_j 表示聚类中心。

本文使用轮廓系数错误!未找到引用源。来评价不同聚类数下的聚类质量，从而找出最佳的聚类数。对于每个聚类的轮廓系数的表示为

$$sil_i=\frac{1}{r_i}\sum_{m=1}^{r_i}\frac{b(m)-a(m)}{\max\{b(m),a(m)\}} \tag{2}$$

其中：r_i 表示每个聚类中样本个数；$a(m)$ 表示样本 m 与同类其他样本的平均距离；$b(m)$ 表示样本 m 与其他聚类中所有样本平均距离的最小值。

对于整个数据集，则可通过平均轮廓指标来评价聚类结果的有效性，表示如下：

$$sil=\frac{1}{N}\sum_{i=1}^{N}sil_i \tag{3}$$

其中：N 表示数据集中样本大小。且 $-1\le sil\le 1$，若 sil 接近 1 时，表示 $a(m)$ 远小于 $b(m)$，即聚类质量效果好。

1.2 经典 GSA 算法

引力搜索算法（GSA）是将空间中所有的粒子视为遵循牛顿第二定律进行无阻力运动的有质量的物体，质量越大的物体将占据更优位置。通过物体间相互万有引力的作用，寻到最优解。经典的 GSA 算法流程如图 2 所示。

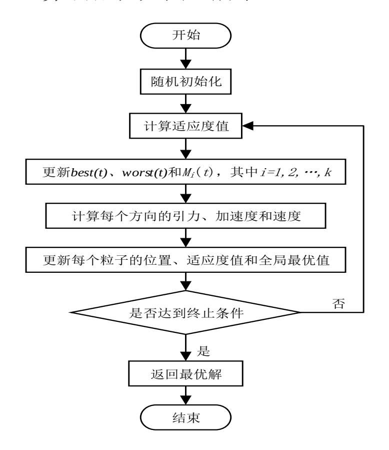


图 2 GSA 算法流程图

经典的引力搜索算法描述如下：假设由 N 个粒子 X_i 构成的种群，在 D 维搜索空间中，定义第 i 个粒子的位置为

$$X_i=(X_i^1,X_i^2,\cdots,X_i^k,\cdots,X_i^D),i=1,2,\cdots,N \tag{4}$$

其中：x_i^k 表示第 i 个粒子在第 k 维上的位置。当进行第 t 次迭代时，粒子 i 的惯性质量 $M_i(t)$ 可根据其适应度值来更新，更新公式为

$$mass_i(t)=\begin{cases}\dfrac{fit_i(t)-worst(t)}{best(t)-worst(t)} & if\ best(t)\ne worst(t)\\ 1 & otherwise\end{cases} \tag{5}$$

$$M_i(t)=\frac{mass_i(t)}{\sum_{j=1}^{N}mass_j(t)} \quad (6)$$

其中：$fit_i(t)$ 表示粒子 i 在迭代 t 时的适应度值，$i=1,2,\cdots,N$。对于求解最小值优化问题，最优适应度 $best(t)$ 和最差适应度 $worst(t)$ 分别为

$$best(t)=\min_{j\in\{1,2,\cdots,N\}} fit_j(t) \quad (7)$$

$$worst(t)=\max_{j\in\{1,2,\cdots,N\}} fit_j(t) \quad (8)$$

反之，即可用于最大值优化问题。

当进行第 t 次迭代时，定义粒子 j 和粒子 i 在第 k 维的相互吸引力为

$$F_{ij}^k(t)=G(t)\frac{M_i(t)\times M_j(t)}{R_{ij}(t)+\varepsilon}(x_j^k(t)-x_i^k(t)) \quad (9)$$

其中：M_j 表示作用粒子 j 的惯性质量；M_i 表示被作用粒子 i 的惯性质量；ε 为常量；$G(t)$ 表示 t 次迭代时的万有引力系数；$R_{ij}(t)$ 表示粒子 i 与粒子 j 之间的距离（一般取欧氏聚类），计算式分别为

$$G(t)=G_0\times e^{(-\alpha\frac{t}{T})} \quad (10)$$

$$R_{ij}(t)=\|X_i(t),X_j(t)\|_2 \quad (11)$$

其中：G_0 表示宇宙在最初时刻的万有引力常量；α 表示引力系数的衰减因子，一般取值为常数；T 表示最大迭代次数。

在 GSA 中，设第 t 次粒子 i 在第 k 维所受的总作用力为

$$F_i^k=\sum_{j=1,j\neq i}^{Kbest} rand_i\times F_{ij}^k(t) \quad (12)$$

其中：$rand_i$ 表示 $[0,1]$ 的一个随机数；F_{ij}^k 的定义如式（6）所示；$Kbest$ 初始值为 N，并随着时间推移逐渐减小为 1，定义为

$$Kbest(t)=final_per+(\frac{1-t}{T})\times(100-final_per) \quad (13)$$

其中：$final_per$ 表示对其他粒子产生作用力的粒子百分比。

依据牛顿第二定律，当进行第 t 次迭代时，粒子 i 在第 k 维上的加速度可定义为

$$a_i^k(t)=\frac{F_i^k(t)}{M_i(t)} \quad (14)$$

GSA 算法每一次迭代进化过程中，粒子根据下式更新粒子 i 的速度 v 和位置 x，即

$$v_i^k(t+1)=rand_j\times v_i^k(t)+a_i^k(t) \quad (15)$$

$$x_i^k(t+1)=x_i^k(t)+v_i^k(t+1) \quad (16)$$

其中：$rand_j$ 表示 $[0,1]$ 的一个随机数。

2　改进的引力搜索算法

本文针对经典引力搜索算法易陷入局部最优解、发生早熟收敛现象等问题，提出一种基于自适应引力系数衰减因子和免疫克隆选择机制的改进引力搜索算法（adaptive attenuation factor based gravitational search algorithm, A2F-GSA）。

2.1　自适应引力系数非线性衰减

在 GSA 算法中，万有引力系数 G 对于寻找算法最优解非常重要，如式（7）所示，其中主要参数为 G_0 和 α，且通常取值为一个常数。研究表明，参数 G_0 通常取值为 100 时，算法能够取得最佳优化效果**错误!未找到引用源。**。对于参数 α，通过调整其取值大小会发现，在算法迭代前期，参数 α 取较小值，能够保证粒子增加的步长，提高算法的全局探索能力；而在算法迭代中后期，当参数 α 取值较大时，将会加快收敛速度，提升算法的局部开发能力**错误!未找到引用源。**。通过研究指数函数的特性，结合参数 α 对算法性能的影响，本文提出一种基于迭代次数自适应变化的引力系数衰减因子，定义如下：

$$\alpha(t)=\gamma\times e^{\lg(\eta\times\frac{t}{T})} \quad (17)$$

其中：t 表示当前迭代次数；T 表示最大迭代次数；γ 和 η 分别为 α 函数的参数。选择合适的参数的来控制 α 函数的变化范围，本文选取 $\gamma=100$，$\eta=0.1$。自适应的 α 在算法早期迭代时生成较大的引力系数 G，能够更加有效的提升全局探索能力，在后期迭代时生成较小的引力系数 G，能够有效提升局部开发能力。

2.2　引入免疫克隆选择机制

通过分析经典 GSA 算法的基本原理可知，随着算法的迭代进化，粒子将逐渐聚集于群体中适应度较优的粒子，从而使得粒子分布收缩，多样性减小，导致算法易陷入局部最优。受文献**错误!未找到引用源。**结合人工蜂群算法和克隆选择算法（Clonal Selection Algorithm，CSA）提出的 DQABCI 算法和文献**错误!未找到引用源。**基于粒子群算法及 CSA 算法提出的 MAPCPSOI 算法启发，本文在经典 GSA 算法中引入免疫克隆选择机制，结合 GSA 算法的寻优能力和 CSA 算法的选择复制、变异和再选择的特点，使得算法具有跳出局部最优的能力，改善算法的早熟收敛现象。

克隆选择算法（CSA）**错误!未找到引用源。**由 De Castro 等人于 2000 年提出，是一种基于人工免疫算法内部微演化过程的优化算法。其以构建记忆单位为基础，实现全局探索与局部开发能力平衡**错误!未找到引用源。**，并由单个最优个体演化为群体最优解集，能够在扩大算法寻优区域的同时，体现免疫系统的多样性**错误!未找到引用源。**。CSA 算法主要包括克隆复制、克隆变异和克隆选择三个阶段，其中克隆复制阶段用于实现种群规模及寻优空间的扩张；克隆变异阶段用于增加种群粒子的多样性，构造出新的种群；克隆选择阶段用于在新建种群中选取适应度（亲和力）高的抗体进入下一代，在实现种群压缩的同时，使得种群朝向更优解移动。文中引入的免疫克隆选择机制具体步骤如下：

（1）粒子亲和度计算

在改进的 GSA 算法中，将粒子视为抗体，因此本文将粒子的适应度函数定义为粒子的亲和度函数，定义如下：

$$afit_i(t) = fit_i(t) \tag{18}$$

其中：$fit_i(t)$ 表示粒子 i 在迭代 t 时的适应度值，$i = 1,2,\cdots,N$ 。

（2）克隆复制

选择当代种群中亲和度值最高的粒子，进行位置信息克隆复制操作，复制个数为 M，文中选取 $M = 10$。

（3）克隆变异

对经过克隆复制产生的新粒子进行位置变异，文中引入基于高斯分布和柯西分布的突变算子，其概率密度函数描述分别如下：

$$f_N(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(x-u)^2}{2\sigma^2}\right] \tag{19}$$

$$f_C(x) = \frac{1}{\pi}\frac{\gamma}{\gamma^2 + (x - x_0)^2} \tag{20}$$

通过分析可知，算法在迭代前期，应保持较大的变异步长，以扩大算法的全局寻优范围；在迭代后期，种群将逐渐聚集于全局最优解，变异步长应逐渐减小，以利于算法收敛。已知基于柯西分布的变异算子相比于高斯变异具有更大的变异尺度，适用于算法前期进化变异；基于高斯分布的变异算子适用于算法的后续迭代，使得算法能够更快地收敛。因此，受文献**错误!未找到引用源。**对于 GSA 算法进行混合变异处理的启发，提出一种改进的随迭代次数自适应调整的变异策略，定义如下：

$$x' = x + \lambda \cdot x \cdot \left[(\frac{t^3}{T^3}) \cdot N(0,1) + (1 - \frac{t^3}{T^3}) \cdot C(0,1) \right] \tag{21}$$

其中：x 表示克隆复制所得粒子位置信息；x' 表示相应粒子变异后的新位置信息；λ 为调节系数，文中取 $\lambda = 0.1$；t 表示算法当前迭代次数；T 表示算法最大迭代次数；$N(0,1)$ 和 $C(0,1)$ 分别表示服从高斯分布与柯西分布的随机数。

（4）克隆选择

经克隆变异之后的变异粒子和原始粒子组成新的粒子种群，表示为 $x'' = [x', x]$。对于新种群 x''，根据粒子的位置信息分别计算其的亲和度值，并从中选取最优个体进入下一次迭代，实现种群压缩的同时，保证解的质量。

3　基于改进 GSA 算法的 K-means 算法

针对 K-means 算法易受聚类中心而陷入局部最优解的问题，本文利用改进引力搜索算法优化聚类中心，且通过调整聚类数的取值得到不同聚类结果，提出了一种基于改进 GSA 算法的 K-means 算法，即 A2F-GSA-Kmeans。

引力搜索算法是通过随机方式寻优不受初始解的影响，因此，本文首先为克服 GSA 算法易陷入局部最优解、发生早熟收敛现象等问题，提出了两种改进策略；然后，将全局搜索能力较强的改进 GSA 算法与 K-means 算法进行结合来优化聚类中心，提出一种新的聚类算法 A2F-GSA-Kmeans 算法；最后，通过利用聚类评价函数对不同聚类结果进行评价，从而获取最佳聚类数。

算法的具体步骤如下：

a)初始化参数。其中包括种群规模 N，算法最大迭代次数 T，随机初始化粒子群体位置 X_i，引力常量 G_0，引力系数衰减因子 α 中各个参数，最优粒子复制数 M 。

b)令算法最初的聚类数 $k = 2$，且 k 的取值范围为 $[2, k_{\max}]$，其中 $k_{\max} \le \sqrt{n}$ 。

c)按照当前指定聚类数对数据集进行聚类，完成个体位置初始化，计算适应度值并选出当前最优解 *Gbest*，并计算相应的聚类有效性指标。

d)根据式（2）（3）更新粒子的惯性质量 $M_i(t)$ 。

e)根据式（7）（14）来更新万有引力系数 $G(t)$ 。

f)根据式（11）计算加速度 a 。

g)根据式（12）（13）更新粒子的速度和位置。

h)克隆复制 M 个当前种群最优粒子，根据式（21）进行变异操作，并从变异新种群中选择最优个体进入下一次迭代。

i)判断是否达到最大迭代次数，如果是，则执行 j)；否则，跳至 c)。

j)令 $k = k+1$，直到 $k > k_{\max}$，转到 Step3。

k)通过比较指标 *sil* 的大小，来寻找最佳的聚类数。

4　实验结果与分析

本文采用 MATLAB R2015b 开发环境进行仿真实验，并在 Windous7 操作系统的计算机上运行，来验证改进算法的有效性。

4.1　改进 GSA 算法性能分析

实验中，为了验证本文提出的 A2F-GSA 改进算法的性能，相关算法参数设置如表 1 所示，并引入 12 个基准测试函数（如表 2 所示）进行仿真实验，其中包括：连续单峰函数（F_1~F_4）、多峰高维函数（F_5~F_8）和多峰低维函数（F_9~F_{12}）。

表 1　算法参数设置

参数	描述	GSA	GG-GSA	A2F-GSA
N	种群规模	50	50	50
T	最大迭代次数	1000	1000	1000
G_0	引力常量初始值	100	100	100
α	引力系数衰减因子	20	30	—
c_1、c_2	学习因子	—	$c_1 = 2 - 1.9(t/T)$ $c_2 = t/T$	—

表 3 展示了 GSA 算法**错误!未找到引用源。**、GG-GSA 算法**错误!未找到引用源。**和 A2F-GSA 算法运行 30 次之后获得的基准测试函数值的平均值、最小值和标准差。图 3 展示了维度为 30 时，GSA 算法与 A2F-GSA 算法优化基准测试函数时收敛过程对比图。图 4 为混合维度下，GSA 算法与 A2F-GSA 算法优化基准测试函数时收敛过程对比图。

表 2　基准测试函数

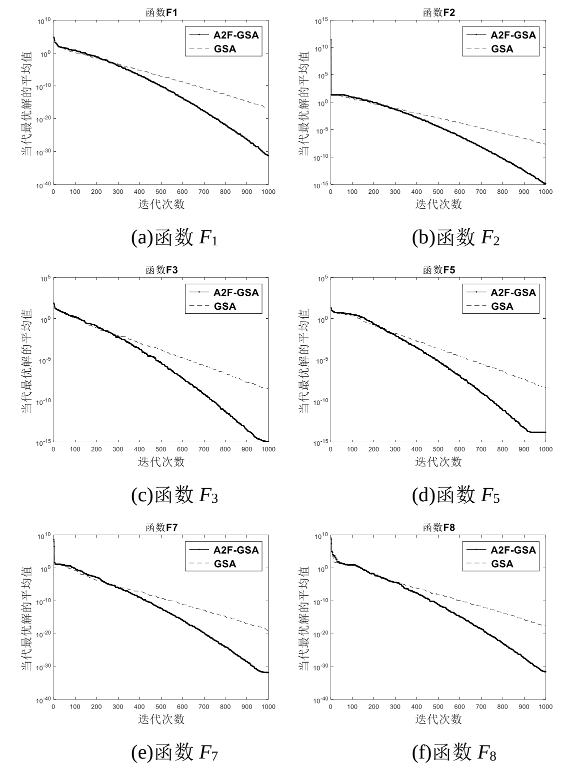
函数名	表达式	维度	范围	理论最优
Schwefel 1.2	$F_1(X) = \sum_{i=1}^{Dim} x_i^2$	30	[-100,100]	0

Schwefel 2.22	$F_2(X)=\sum_{i=1}^{Dim}\lvert x_i\rvert+\prod_{i=1}^{Dim}\lvert x_i\rvert$	30	[-10,10]	0
Schwefel 2.21	$F_3(X)=\max_i\{\lvert x_i\rvert,1\le i\le D\}$	30	[-100,100]	0
Step	$F_4(X)=\sum_{i=1}^{Dim}\left(\lfloor x_i+0.5\rfloor\right)^2$	30	[-100,100]	0
Ackley	$F_5(X)=-20\exp\left(-0.2\sqrt{\frac{1}{Dim}\sum_{i=1}^{Dim}x_i^2}\right)-\exp\left(\frac{1}{Dim}\sum_{i=1}^{Dim}\cos(2\pi x_i)\right)+20+e$	30	[-32,32]	0
Griewank	$F_6(X)=\frac{1}{4000}\sum_{i=1}^{Dim}x_i^2-\prod_{i=1}^{Dim}\cos\left(\frac{x_i}{\sqrt{i}}\right)+1$	30	[-600,600]	0
Penalized	$F_7(X)=\frac{\pi}{Dim}\left\{\sin^2(3\pi y_1)+\sum_{i=1}^{Dim-1}(y_i-1)^2\left[1+\sin^2(3\pi y_i+1)\right]+(y_{Dim}-1)^2\right\}+\sum_{i=1}^{Dim}u(x_i,10,100,4)$	30	[-50,50]	0
Generalized Penalized	$F_8(X)=0.1\left\{\sin^2(3\pi x_1)+\sum_{i=1}^{Dim}(x_i-1)^2\left[1+\sin^2(3\pi x_i+1)\right]+(x_{Dim}-1)^2\left[1+\sin^2(3\pi x_{Dim})\right]\right\}+\sum_{i=1}^{Dim}u(x_i,5,100,4)$	30	[-50,50]	0
Shekel	$F_9(X)=\left[\frac{1}{500}+\sum_{j=1}^{25}\frac{1}{j+\sum_{i=1}^{2}(x_i-a_{ij})^6}\right]^{-1}$	2	[-65.536,65 .536]	1
Goldstein-Pri ce	$F_{10}(X)=\left[1+(x_1+x_2+1)^2(19-14x_1+3x_1^2-14x_2+6x_1x_2+3x_2^2)\right]\times\left[30+(2x_1-3x_2)^2\times(18-32x_1+12x_1^2+48x_2-36x_1x_2+27x_2^2)\right]$	2	[-2,2]	3
Hartman	$F_{11}(X)=\sum_{i=1}^{4}\exp[-\sum_{j=1}^{3}a_{ij}(x_j-p_{ij})^2]$	3	[0,1]	-3.86
Shekel's Family	$F_{12}(X)=-\sum_{i=1}^{10}\left[(x-a_i)(x-a_i)^T+c_i\right]$	4	[0,10]	−10.5 363

表 3　A2F-GSA 与其他算法优化基准函数对比

函数		GSA 错误!未找到引用源。	GG-GSA 错误!未找到引用源。	A2F-GSA
F_1	*Mean*	2.244E - 17	5.417E - 23	**7.032E - 32**
	Best	1.130E - 17	2.896E - 25	**5.257E - 32**
	Std.Dev	6.862E - 18	2.201E - 22	**3.017E - 32**
F_2	*Mean*	2.289E - 08	1.260E - 10	**1.442E - 15**
	Best	1.802E - 08	3.510E - 12	**1.282E - 15**
	Std.Dev	2.658E - 09	2.337E - 10	**1.643E - 16**
F_3	*Mean*	3.567E - 02	1.741E - 09	**3.443E - 15**
	Best	2.488E - 09	1.232E - 10	**9.360E - 16**
	Std.Dev	1.318E - 01	1.877E - 09	**3.004E - 15**
F_4	*Mean*	0.000E + 00	0.000E + 00	**0.000E + 00**
	Best	0.000E + 00	0.000E + 00	**0.000E + 00**
	Std.Dev	0.000E + 00	0.000E + 00	**0.000E + 00**
F_5	*Mean*	3.597E - 09	2.703E - 12	**1.013E - 14**
	Best	2.407E - 09	4.343E - 13	**7.994E - 15**
	Std.Dev	6.866E - 10	3.218E - 12	**1.946E - 15**
F_6	*Mean*	4.343E + 00	1.170E + 00	**4.045E - 02**
	Best	1.985E + 00	3.946E - 02	**0.000E + 00**
	Std.Dev	1.735E + 00	8.023E - 01	**4.322E - 02**
F_7	*Mean*	2.015E - 02	3.455E - 02	**1.041E - 02**
	Best	7.913E - 20	1.277E - 27	**1.571E - 32**
	Std.Dev	4.799E - 02	9.940E - 02	**3.162E - 02**
F_8	*Mean*	1.590E - 03	7.325E - 04	**3.663E - 04**
	Best	1.297E - 18	2.838E - 26	**1.350E - 32**
	Std.Dev	5.395E - 03	2.788E - 03	**2.006E - 03**
F_9	*Mean*	3.726E + 00	**2.241E + 00**	2.247E + 00
	Best	9.980E - 01	9.980E - 01	**9.980E - 01**
	Std.Dev	2.726E + 00	**9.922E - 01**	1.327E + 00
F_{10}	*Mean*	3.000E + 00	3.000E + 00	**3.000E + 00**
	Best	3.000E + 00	3.000E + 00	**3.000E + 00**
	Std.Dev	4.807E - 15	2.566E - 15	**5.470E - 16**
F_{11}	*Mean*	-3.863E + 00	-3.863E + 00	**-3.863E + 00**
	Best	-3.863E + 00	-3.863E + 00	**-3.863E + 00**
	Std.Dev	2.710E - 15	2.710E - 15	**2.220E - 16**
F_{12}	*Mean*	-1.029E + 01	-1.054E + 01	**-1.054E + 01**
	Best	-1.054E + 01	-1.054E + 01	**-1.054E + 01**
	Std.Dev	1.323E + 00	9.034E - 15	**1.546E - 15**

从表 3 可以看出，在 12 个基准测试函数中，针对单峰函数的优化求解，A2F-GSA 算法对于函数 F_1-F_3 的收敛精度均明显优于其他算法，而对于步进函数 F_4，经典 GSA 及其改进算法均可取得其理论最优解，且由图 3(a)~(c)可知，收敛速度相对于经典的 GSA 算法显著提高。表明本文提出的自适应引力系数衰减方法能够有效平衡算法的全局探索能力和局部开发能力，从而有效提升算法的求解效果与效率。

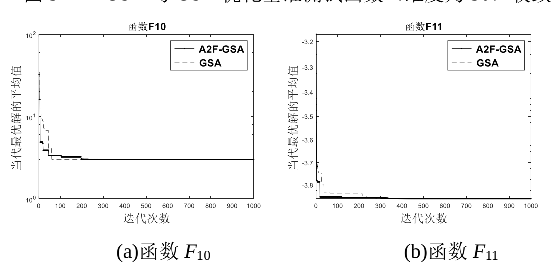


(a)函数 F_1　(b)函数 F_2

(c)函数 F_3　(d)函数 F_5

(e)函数 F_7　(f)函数 F_8

图 3 A2F-GSA 与 GSA 优化基准测试函数（维度为 30）收敛曲线



(a)函数 F_{10}　(b)函数 F_{11}

图 4 A2F-GSA 与 GSA 优化基准测试函数（混合维度）收敛曲线

针对具有多个局部极值的多峰函数的优化求解，从表 3 和图 3(d)~(f)可知，在维度为 30 的测试条件下，A2F-GSA 算法对于函数 F_5-F_8 的优化精度与速度显著优于其他算法，表明将免疫克隆选择机制引入 A2F-GSA 算法中，能够使得算法有效跳出局部最优，改善算法早熟收敛现象。而在混合维度下，函数 F_9-F_{12} 相比于高维多峰函数具有较少的局部极值点，由表 3 和图 4(a)(b)可知，A2F-GSA 算法对于 F11、F_{12}、F_{13} 的优化求解均取得相应的理论最优解，且标准差值均低于其他算法，只是对于函数 F_9 的优化效果略劣于 GG-GSA 算法，但相比于经典 GSA 算法仍有显著提高。

综上所述，对于文中选取的 12 个基准测试函数，本文提出

的 A2F-GSA 算法具有最优的求解精度、收敛速度和鲁棒性。

4.2 改进聚类算法性能分析

为了验证本文提出的基于改进的引力搜索算法的 K-means 算法的有效性，从而寻找最佳聚类数，本文采用公开测试数据集 UCI 错误!未找到引用源。库中的数据集 Normal07、Cancer、Iris、Wine、Glass、Abalone 进行实验仿真。且数据集的特征分布如表 4 所示。

表 4 数据集特征描述

数据集	大小	特征数	类别数
Normal07	7000	2	7
Cancer	569	30	2
Iris	150	4	3
Wine	178	13	3
Glass	214	9	6
Abalone	4177	7	23

算法中的参数设置采用 A2F-GSA 算法的基本参数设置，聚类数的搜索范围为 $[2, k_{max}]$，其中 $k_{max} \le \sqrt{n}$。实验中，使用本文提出的 A2F-GSA-Kmeans 算法和经典的 GSA 算法对所选取的测试数据集各自运行 20 次，分别记录 20 次中得到正确聚类结果的次数，通过计算得到算法的聚类准确率，计算公式如下：

$$\text{聚类准确率（\%）} = \frac{\text{得到正确聚类结果的次数}}{\text{实验运行总次数}} \times 100\% \qquad （22）$$

并与文献**错误!未找到引用源。**中基于改进蜂群算法的 K-means 算法进行比较，实验结果如表 5 所示。

表 5 算法准确率对比

算法	聚类准确率%					
	Normal07	Cancer	Iris	Wine	Glass	Abalone
GSA-Kmeans	100	50	90	70	40	30
Improved-ABC-Kmean 错误!未找到引用源。	—	—	90	—	50	—
A2F-GSA-Kmeans	100	95	100	95	90	70

由表 5 可知，对于数据维度低且分布明显分离的 Normal07，每次运行均能得到正确的聚类数；对于数据分离不明显的 Iris 和 Wine，A2F-GSA-Kmeans 算法得到聚类准确率略优于经典的 GSA 聚类算法。而对于复杂数据集 Glass、高维数据集 Cancer，A2F-GSA-Kmeans 算法的聚类准确率明显优于其他算法，从而验证了本文改进算法的有效性。然而，对于带有噪声的数据集 Abalone，均未能取得较好的聚类准确率，有待进一步的研究。

5 结束语

本文首先提出一种自适应引力系数衰减因子函数来代替常量值，使得引力系数 *G* 从大到小非线性变化，有效提高算法的全局探索能力和局部开发能力；同时，将免疫克隆选择机制引入 GSA 算法中，能够使算法有效跳出局部最优，且改善算法早熟收敛现象。在 12 个基准测试函数上的仿真实验结果，验证了本文所提出的 A2F-GSA 算法相比于其他算法具有更好的优化收敛性能。然后结合 A2F-GSA 算法和 K-means 算法提出一种新的 A2F-GSA-Kmeans 聚类算法。并在 6 个 UCI 测试数据集上进行实验表明，相比于基于经典 GSA 和改进蜂群算法的 K-means 算法，本文提出的 A2F-GSA-Kmeans 算法的聚类质量有明显提高。

参考文献：

[1] Chen T W, Ikeda M. Design and implementation of low-power hardware architecture with single-cycle divider for on-line clustering algorithm [J]. IEEE Trans on Circuits & Systems I Regular Papers, 2013, 60 (8): 2165-2176.

[2] 杨菊婧，张达敏. 基于改进 BA 算法的 K-means 聚类 [J]. 计算机应用研究，2018，35 (05): 1454-1457. (Yang Juqing, Zhang Damin. K-means clustering algorithm based on improved BA algorithm [J]. Application Research of Computers, 2018, 35 (05): 1454-1457.)

[3] 于佐军，秦欢. 基于改进蜂群算法的 K-means 算法 [J]. 控制与决策，2018, 33 (1): 181-185. (Yu Zuojun, Qin Huan. K-means algorithm based on improved artificial bee colony algorithm [J]. Control and Decision, 2018, 33 (1): 181-185.)

[4] Rashedi E, Nezamabadi-Pour H, Saryazdi S. GSA: a gravitational search algorithm [J]. Information Sciences. 2009, 179 (13): 2232-2248.

[5] Liu Chao, Niu Peifeng, Li Guoqiang, *et al.* A hybrid heat rate forecasting model using optimized LSSVM based on improved GSA [J]. Neural Processing Letters, 2017, 45 (1): 299-318.

[6] Sun Genyun, Zhang Aizhu, Wang Zhenjie, *et al.* Locally informed gravitational search algorithm [J]. Knowledge-Based Systems, 2016, 104 (C): 134-144.

[7] Mirjalili S, Gandomi A H. Chaotic gravitational constants for the gravitational search algorithm [J]. Applied Soft Computing, 2017, 53: 407-419.

[8] Mirjalili S, Lewis A. Adaptive gbest-guided gravitational search algorithm [J]. Neural Computing & Applications, 2014, 25 (7-8): 1569-1584.

[9] Darzi S, Kiong T S, Islam M T, *et al.* A memory-based gravitational search algorithm for enhancing minimum variance distortionless response beamforming [J]. Applied Soft Computing, 2016, 47 (C): 103-118.

[10] Vijay K B, K. V. Arya, An effective gbest-guided gravitational search algorithm for real-parameter optimization and its application in training of feedforward neural networks [J]. Knowledge-Based Systems, 2018, 143: 192-207.

[11] 朱连江，马炳先，赵学泉. 基于轮廓系数的聚类有效性分析 [J]. 计算机应用, 2010, 30 (S2): 139-141+198. (Zhu Lianjiang, Ma Bingxian, Zhao Xuequan. Clustering validity analysis based on silhouette coefficient [J]. Journal of Computer Applications, 2010, 30 (S2): 139-141+198.)

[12] 范炜锋. 万有引力搜索算法的分析与改进 [D]. 广州：广东工业大学，2014. (Fang Weifeng. Analysis and improvement of gravitational search algorithm [D]. Guangzhou: Guangdong University of Technology, 2014.)

[13] 蒋建国，谭雅，董立明，等. 改进的万有引力搜索算法在边坡稳定分析中的应用 [J]. 岩土工程学报, 2016, 38 (3): 419-425. (Jiang Jianguo, Tan Ya, Dong Liming, *et al.* Application of modified gravitational search algorithm in slope stability analysis [J]. Chinese Journal of Geotechnical Engineering, 2016, 38 (3): 419-425.)

[14] 赵辉，李牧东，翁兴伟. 分布式人工蜂群免疫算法求解函数优化问题 [J]. 控制与决策, 2015, 30 (7): 1181-1188. (Zhao Hui, Li Mudong, Weng Xingwei. Distributed artificial bee colony immune algorithm for the problems of function optimization [J]. Control and Decision, 2015, 30 (7): 1181-1188.)

[15] 吴建辉，章兢，李仁发，等. 多子种群微粒群免疫算法及其在函数优化中应用 [J]. 计算机研究与发展, 2012, 49 (9): 1883-1898. (Wu Jianhui, Zhang Jing, Li Renfa, *et al.* A multi-subpopulation PSO immune algorithm and its application on function optimization [J]. Journal of Computer Research and Development, 2012, 49 (9): 1883-1898.)

[16] DeCastro L N, Zuben V. Learning and optimization using the clonal seletion. Issue on Artificial Immune System (AIS) . 2002, 6 (3): 239-351.

[17] 舒万能，丁立新. 克隆选择算法的优化和品质因数 [J]. 软件学报, 2016, 27 (11): 2763-2776. (Shu Wanneng, Ding Lixin. Optimization and quality factor of clonal selection algorithm [J]. Journal of Software, 2016, 27 (11): 2763-2776.)

[18] Mohammadi M, Raahemi B, Akbari A, *et al.* Improving linear discriminant analysis with artificial immune system-based evolutionary algorithms [J]. Information Sciences, 2012, 189 (7): 219-232.

[19] Zhang Nan, Li Chaoshun, Li Ruhai, *et al.* A mixed-strategy based gravitational search algorithm for parameter identification of hydraulic turbine governing system [J]. Knowledge-Based Systems, 2016, 109: 218-237.

[20] University of California, Irvine. UCI machine learning repository [DB/OL]. [2013-06-19]. http://archive. ics. uci. edu/ml/datasets. html